line 1 - B

MANE 3332.01

LECTURE 18

Agenda

- Midterm exams are not graded; still contacting students who missed
- Linear Combination Practice Problems (assigned 10/28, due 10/30)
- Linea Combination Quiz (assigned 10/30, due 11/4)
- Complete Chapter Six and Start Chapter 7
- Attendance
- Questions?

Handouts

- Lecture 18 Slides
- Lecture 18 Slides - marked

CHAPTER 6, CONTINUED

Calculating Quantiles

Find x such that

$P(X \le x) = p$

Let $p = .10$

x_p or $x_{.10}$

Quartiles

$p = .25$ or $.5$ or $.75$

Deciles

$p = .1$ or $.2$ or $\cdots$ $.9$

Percentiles

$p = .01, .02, \ldots, .99$

2.3.2 Sample Quantiles

In Example 2.8, we consider an ogive for the plated bracket data. The point (1.55, 0.567) is on that ogive, so we estimate that 56.7% of the sampled population of brackets weighed at most 1.55 ounces. Weights associated with other percentages can also be estimated by locating the appropriate point on the ogive. In general, if the point (x, p) is on the ogive, we can use x as an estimate of the weight with $100p$% of the population values at or below it. This estimate, called the 100pth *sample quantile*, is denoted x_p.

If two persons (or computer programs) use different groupings to obtain an ogive, the resulting quantiles will differ. To remedy this deficiency, an algebraic procedure is required.

THE 100pth SAMPLE QUANTILE

Several definitions of sample quantiles are used. We use the one that agrees with the default values output by the UNIVARIATE procedure in SAS®. Also, the definition used here is consistent with our definition of the sample median.

Suppose a sample of size n is obtained from some population associated with a continuous variable. For $0 < p < 1$, let $p(n+1) = i + d$, with i the integer part of $p(n+1)$ and $0 \le d < 1$ the decimal part. If $1 \le i < n$, and $d = 0$, the 100pth sample quantile is $x_{(i)}$. If $1 \le i < n$ and $0 < d < 1$, interpolate linearly between $x_{(i)}$ and $x_{(i+1)}$. In either case, the 100pth *sample quantile* is

$$x_p = x_{(i)} + d\,[x_{(i+1)} - x_{(i)}] \tag{2.4}$$

when $1 \le i < n$. If $i = 0$ or n, the 100pth sample quantile does not exist. If $100p$ is an integer, the corresponding quantile is called a *percentile*.

EXAMPLE 2.18

Suppose we want to find the 43rd percentile of the sample of plated weights in Table 2.1. Since there are $n = 75$ observations in the sample and $p = 0.43$, we find $p(n+1) = (0.43)(75+1) = 32.68$. Letting $i = 32$ and $d = 0.68$, we use Equation (2.4) to obtain $x_{0.43} = x_{(32)} + (0.68)(x_{(33)} - x_{(32)})$. The 32nd ordered value in Figure 2.1(b) is $x_{(32)} = 1.50$ and the 33rd ordered value is $x_{(33)} = 1.51$. Thus, the 43rd percentile for these data is $x_{0.43} = 1.50 + (0.68)(1.51 - 1.50) = 1.5068 \approx 1.507$. Using this as a point estimate of the population percentile, we can say that approximately 43% of the plated brackets produced on the day the data were collected had weights of 1.507 ounces or less. ■

The Sample Median Is a Percentile

Suppose we want to find the 50th percentile and the data set contains n values. When n is even, $(0.50)(n+1) = (n/2) + (0.50)$, with $n/2$ a positive integer. Using Equation (2.4) with $i = n/2$ and $d = 0.50$, $x_{0.50} = x_{(i)} + (0.50)[x_{(i+1)} - x_{(i)}] = [x_{(i)} + x_{(i+1)}]/2$. When n is odd, $(0.50)(n+1) = (n+1)/2$, with $(n+1)/2$ a positive integer. Using Equation (2.4) with $i = (n+1)/2$ and $d = 0$, we find $x_{0.50} = x_{(i)}$. But, this is precisely how the sample median was defined. Thus, $\tilde{x} = x_{0.50}$.

SAMPLE QUARTILES

The percentiles $x_{0.25}$, $x_{0.50}$, and $x_{0.75}$ are known as the *first*, *second*, and *third sample quartiles*, respectively. These quantities are often denoted q_1, q_2, and q_3.

EXAMPLE 2.19

Consider the plated bracket weights in Table 2.1. Using the ordered stem-and-leaf display presented in Figure 2.1(b), we find the following.

(a) *First Quartile:* Since $(0.25)(75+1) = 19$, $q_1 = x_{0.25} = x_{(19)} = 1.46$.

(b) *Second Quartile (Median):* Since $(0.50)(75+1) = 38$, $q_2 = \tilde{x} = x_{0.50} = x_{(38)} = 1.53$.

1, 2, 3, 4

median

$p = .5$

$.5(4+1) = 2.5$

1 $p(n+1) = i + d$

$i = 2$, $d = .5$

2 $x_p = x_{(i)} + d[x_{(i+1)} - x_{(i)}]$

reference for calculating quantiles

Quantiles

Quantile Example

$x_{(i)}$ - ith ranked observations

8 Observations from binomial distribution with $n = 10$ ~~and $p = .5$~~

	1	2	3	4	5	6	7	8
x_i	6,	4,	5,	7,	3,	5,	4,	6
$x_{(i)}$	3	4	4	5	5	6	6	7

$x_2 = 4$, $x_{(2)} = 4$, $x_3 = 5$, $x_{(3)} = 4$

$$x_p = x_{(i)} + d\left[x_{(i+1)} - x_{(i)}\right]$$

Quantile Example

$p = .75$

$p(n+1) = .75(8+1) = \underline{6.75}$

$i = 6$ & $d = .75$

$$x_{.75} = x_{(6)} + .75\left[x_{(7)} - x_{(6)}\right]$$

$$= 6 + .75(6 - 6)$$

Exploratory Data (Graphical) Analysis

- Exploratory data analysis (EDA) is the use of graphical procedures to analyze data.
- John Tukey was a pioneer in this field and invented several of the procedures
- Tools include stem-and-leaf diagrams, box plots, time series plots and digidot plots

Stem and Leaf Diagram

- Excellent tool that maintains data integrity
- The stem is the leading digit or digits
- The leaf is the remaining digit
- Make sure to include units
- R Code

```
stem(midterm$MidtermExam)
```

leaf unit

Stem and Leaf Example

R output of a Stem and Leaf diagram

stem leaf

```
The decimal point is 1 digit(s) to the right of the |

2 | 8
3 |
4 | 4
5 | 11566
6 | 1333444667 9
7 | 2247
8 | 00147
9 | 2
```

Stem = 2
leaf = 8

28.0

Stem and Leaf Plot of Midterm Exam Scores

Histogram

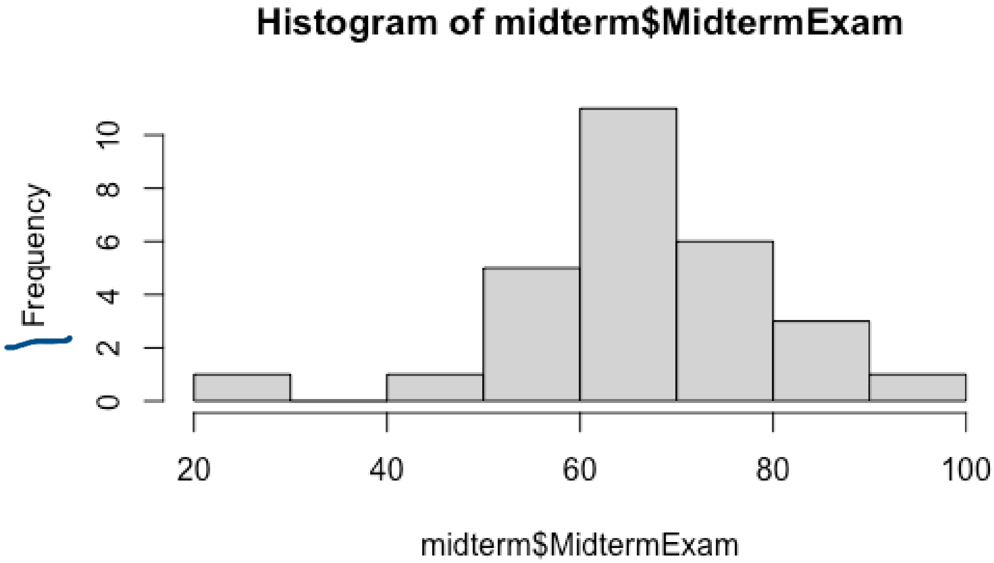
- A histogram is a barchart displaying the frequency distribution information
- There are three types of histograms: frequency, relative frequency and cumulative relative frequency
- R code

```
hist(midterm$MidtermExam)
```

Histogram Example

R output of histogram

Relative Freq



Histogram of Midterm Exam Scores

Box & whiskers

IQR - inter quartile range
$= Q_3 - Q_1$

Boxplot

Graphical display that simultaneously describes several important features of a data set such as center, spread, departure from symmetry and outliers

Requires the calculation of quantiles (quartiles)

Box Plot 1

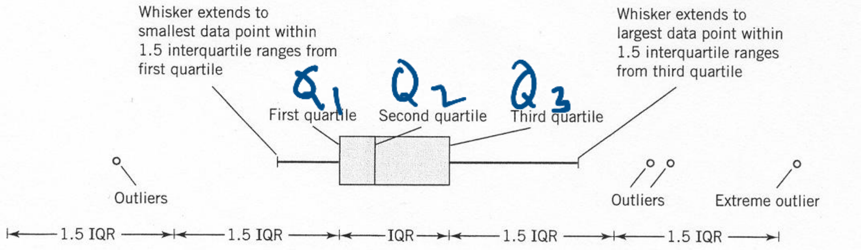


Figure 2-11
Description of a box plot.

Potential outliers

Box plot with explanation

Box Plot 2

Highest Quality level

Most Variability

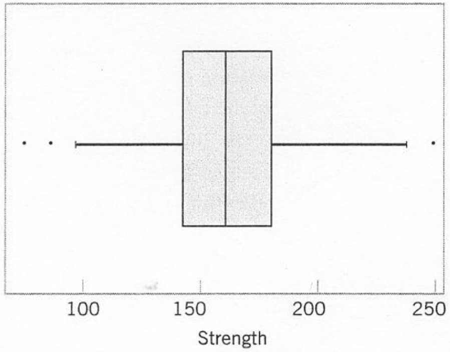


Figure 2-12 Box plot for compressive strength data in Table 2-2.

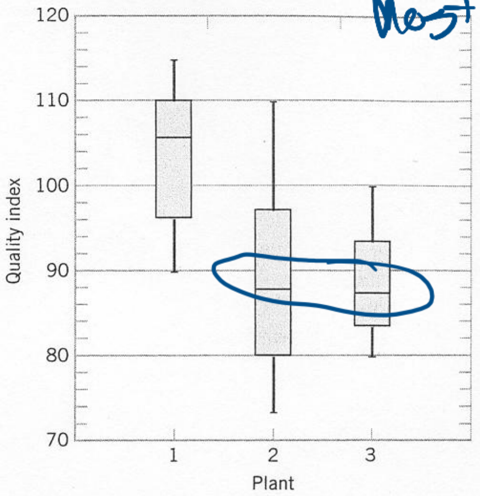


Figure 2-13 Comparative box plots of a quality index at three plants.

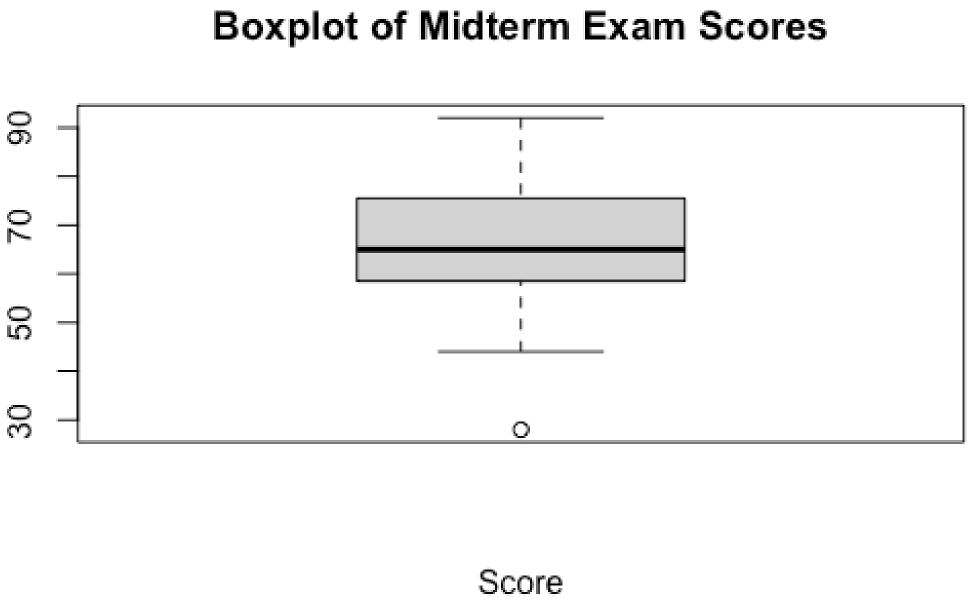
examples of boxplots

Box Plot 3

R code for Box Plot

```
boxplot(midterm$MidtermExam,xlab='Score',main='Boxplot of Midterm Exam Scores')
```

R Box Plot output



Boxplot of Midterm Exam Scores

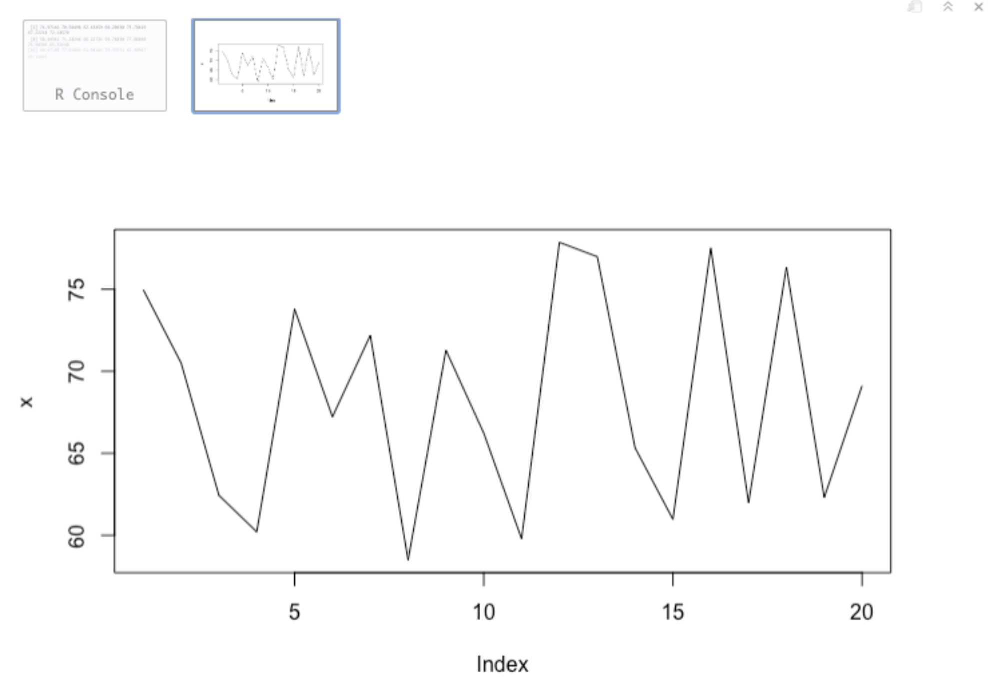
Time Series Plot

- A **time series plot** is a graph in which the vertical axis denotes the observed value of the variable (say x) and the horizontal axis denotes time
- Excellent tool for detecting:
  - trends,
  - cycles,
  - other non-random patterns

Time Series Plot in R

```
```{r}
x<-rnorm(20,70,10)
print(x)
plot(x,type='l')
```
```

R Code



Time Series Plot

Probability Plotting

- **Probability plotting** is a graphical method of determining whether sample data conform to a hypothesized distribution
- Used for validating assumptions
- Alternative to hypothesis testing

Construction

1. Sort the data from smallest to largest, .
2. $x_{(1)}, x_{(2)}, \dots, x_{(n)}$
3. Calculate the observed cumulative frequency $(j - 0.5)/n$

For the normal distribution find z_j that satisfies

$$\frac{j - 0.5}{n} = P(Z \leq z_j) = \Phi(z_j)$$

3. Plot z_j versus $x_{(j)}$ on special graph paper

Usage

If the data plots as a straight line, the assumed distribution is correct

6-7 PROBABILITY PLOTS 215

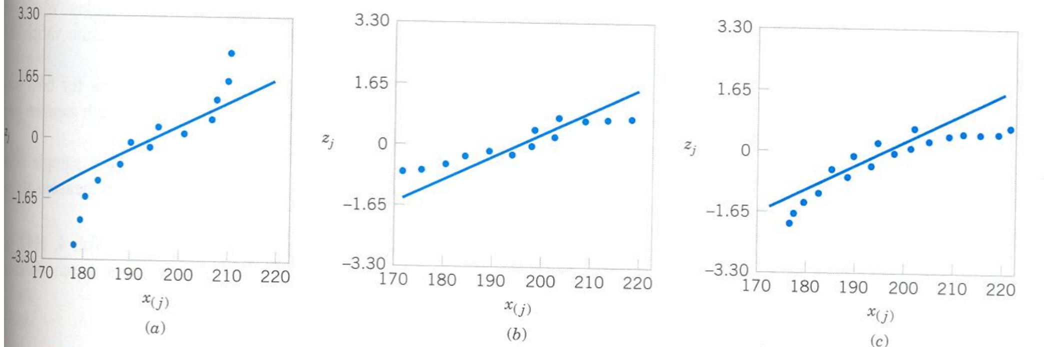
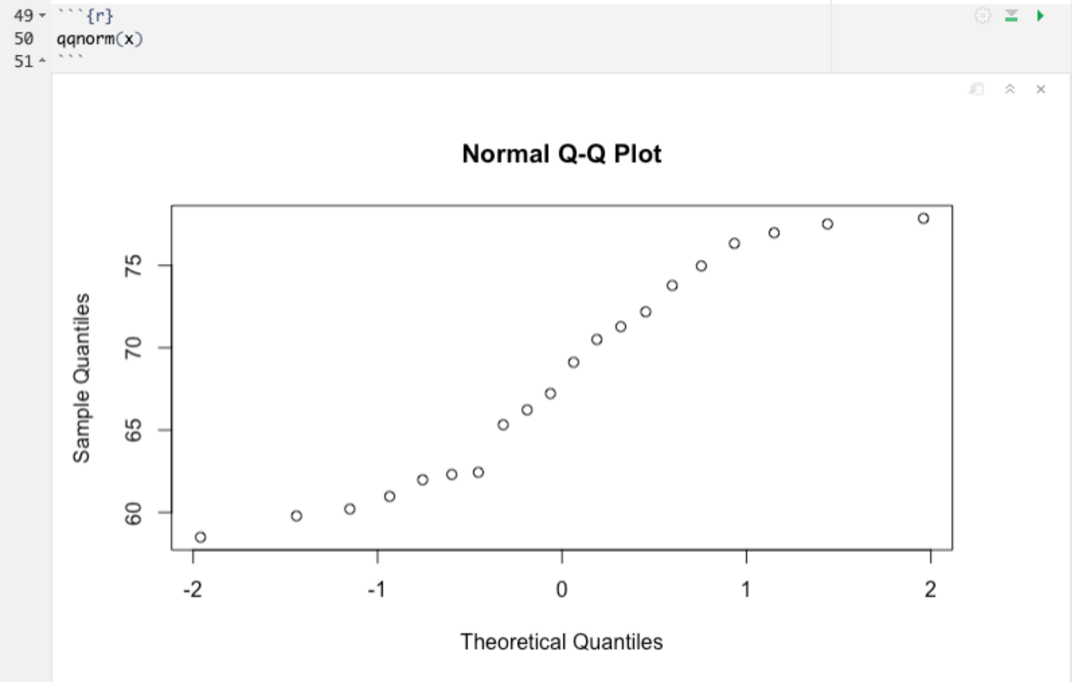


Figure 6-21 Normal probability plots indicating a nonnormal distrubution. (a) Light-tailed distribution. (b) Heavy-tailed distribution. (c) A distribution with positive (or right) skew.

normal probability plots from textbook, figure 6.21 on page 215

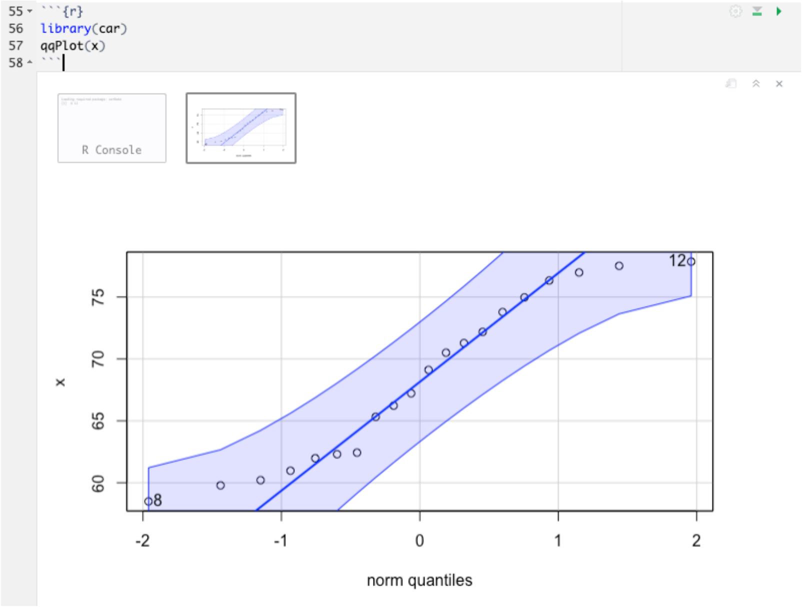
Probability Plot Example 1 in R



Normal Probability Plot

Probability Plot Example 2

- Difficulty from example one is how close to straight is “good enough”
- Add confidence bands to normal probability plot
  - Requires package car to be added to R
  - If all points are within the band, we are 95% confident that the sample is from a normal distribution. However if one or more points are not within band, the data is not from a normal distribution

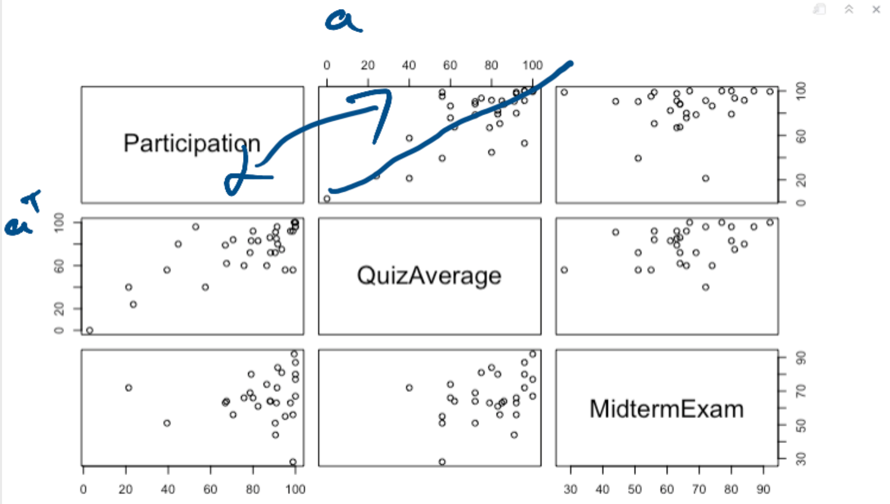


QQ Plot with band

MULTIVARIATE DATA

Matrix of Scatter Plot in R

```{r}
plot(midterm)
```



Scatter Plots

Covariance in R

```{r}
midterm_NA <- na.omit(midterm)
print(cov(midterm_NA))
```

```
              Participation QuizAverage MidtermExam
Participation     340.16778    193.7847    28.75699
QuizAverage       193.78474    269.0899    81.17460
MidtermExam        28.75699     81.1746   188.43915
```

Covariance Matrix

Correlation

```{r}
print(cor(midterm_NA))
```

```
              Participation QuizAverage MidtermExam
Participation     1.0000000   0.6405076   0.1135825
QuizAverage       0.6405076   1.0000000   0.3604839
MidtermExam       0.1135825   0.3604839   1.0000000
```

Correlation Matrix

Chapter 7 Overview

- Chapter 7 contains a detailed explanation of point estimates for parameters
- Much of this chapter is of a highly statistical nature and will not be covered in this course
- Key concepts we will discuss are:
  - Statistical inference
  - Statistic
  - Sampling distribution
  - Point estimator
  - Unbiased estimate
  - MVUE estimator
  - Central limit theorem
  - Sampling distributions

Statistical Inference

- Montgomery gives the following description of statistical inference.

> The field of statistical inference consists of those methods used to make decisions or to draw conclusions about a population. There methods utilize the information contained in a sample from the population in drawing conclusions. This chapter begins our study of the statistical methods used for inference and decision making.

- Statistical inference may be divided into two major areas: parameter estimation and hypothesis testing

Chapter 8 Chapter 9 & 11

Point Estimate

- Montgomery states that "In practice, the engineer will use sample data to compute a number that is in some sense a reasonable value (or guess) of the true mean. This number is called a **point estimate**."
- Discuss examples
- A formal definition of a point estimate is

> A **point estimate** of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\widehat{\Theta}$. The statistic $\widehat{\Theta}$ is called the point estimate.

- Notice the use of the "hat" notation to denote a point estimate

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = f(X_1, X_2, \dots, X_n)$$

Statistic

$\bar{X} = \frac{1}{n} \sum_{i=1}^{n}$ as $n \to \infty$ $\bar{X}$ has a normal distribution

- Point estimate requires a sample of random observations, say $X_1, X_2, \ldots, X_n$
- Any function of the sampled random variables is called a statistic
- The function of the random variables is itself a random variable
- Thus, the sample mean $\bar{x}$ and the sample variance s^2 are both statistics and random variables

Properties of point estimators

- We would like point estimates to be both accurate and precise
- An unbiased estimator addresses the accuracy criteria
- A minimum variance unbiased estimator addresses the precision criteria

Unbiased Estimator

- The point estimator $\widehat{\Theta}$ is an **unbiased estimator** for the parameter θ if

$$E\left(\widehat{\Theta}\right) = \theta$$

- If the point estimator is not unbiased, then the difference

$$E\left(\widehat{\Theta}\right) - \theta$$

is called the **bias** of the estimator $\widehat{\Theta}$

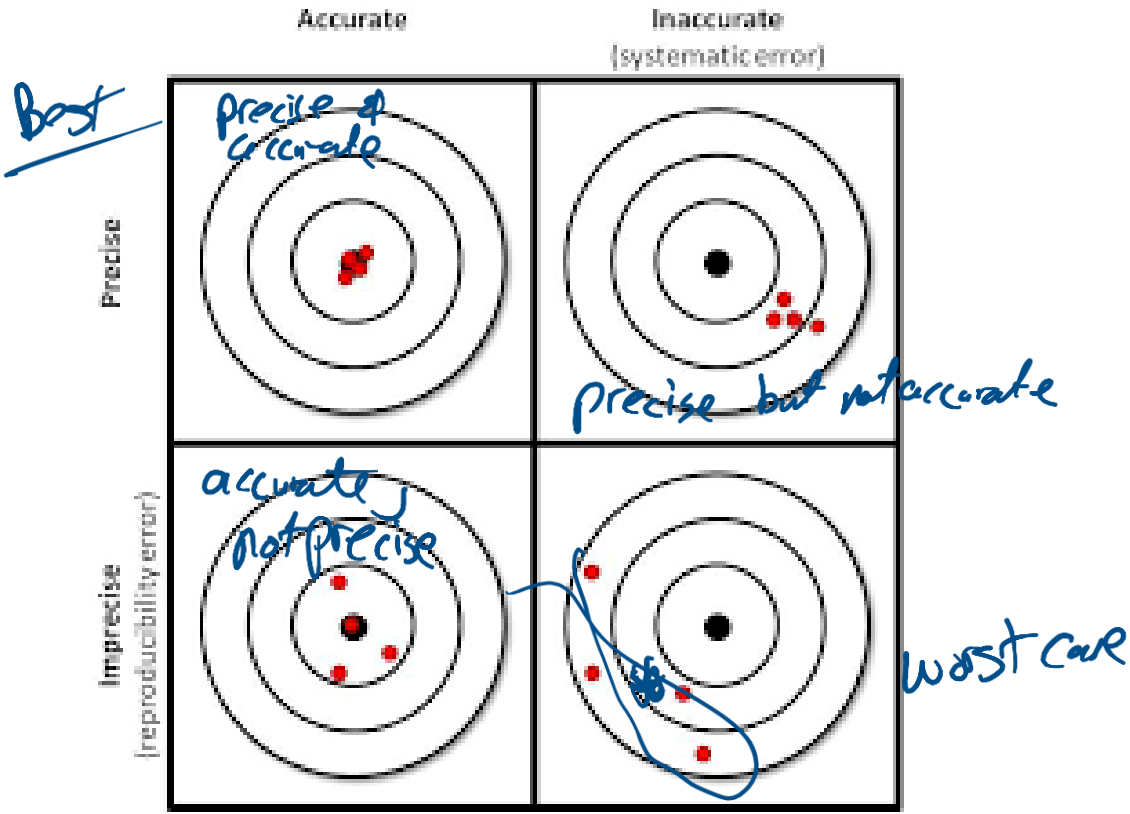
MVUE

- Montgomery gives the following definition of a minimum variance unbiased estimator (MVUE)

  > If we consider all unbiased estimators of θ, the one with the smallest variance is called the minimum variance unbiased estimator

- An import fact is that the sample mean $\bar{x}$ is the MVUE for μ when the data comes from a normal distribution

Accuracy vs. Precision



graph of accuracy vs. precision

Sampling Distribution

- The probability distribution of a statistic is called a **sampling distribution**

Central Limit Theorem

- Definition of the Central Limit Theorem is

    If $X_1, X_2, \dots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2, and if $\overline{X}$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

as $n \to \infty$, is the standard normal distribution

- Important result because for sufficiently large n, the sampling distribution of $\overline{X}$ is normally distribution
- This is a fundamental result that will be used extensively in the next four chapters of the textbook.